

Partial Translation of JP 1983-193595

Publication Date: November 11, 1983

Application No.: 1982-75282

Filing Date: May 7, 1982

Applicant: HITACHI LTD

Inventor: Kazuo NAKADA

Line 8 of right top column, page 3 to line 4 of left top column, page 5

The point of the present invention is an attempt to enable voice and PB signals to be used together as an information input means, by adding to the voice recognition unit 13, as already described in FIG. 1, a PB signal, as a phoneme and one word, and detecting the PB signal in an exactly same format as voice recognition. However, the invention does not assume simultaneous coexistence of a voice signal and a PB signal.

First, we simply consider the case in which frame-by-frame phoneme recognition takes place for each and every one of 16 pairs of phoneme patterns. In this case, one each of 16 kinds of PB signals, in total, is allocated as a pseudo-phoneme norm to each of the 16 pairs, and feature patterns necessary for its detection may be stored in the phoneme norm pattern memory. A word dictionary associated with PB signals may be configured to satisfy the condition that a pseudo-phoneme norm for the same PB signal be maintained for a time longer than the time in which reception and detection should take place (for instance, 40 milliseconds or longer according to the prevailing regulations).

Then, we consider the following in the case of hierarchical process in which a first column is recognized by first 2 representative clusters:

- 1) It should be detected through recognition of a first column that it is a PB signal; and
- 2) If it has been detected as PB signal in recognition of the first column, then, on a second column, recognize which one of them it is.

In the following, we describe more specifically.

Now we consider the case in which in phoneme recognition, matching according to a likelihood ratio is taken based on LPC (linear prediction) analysis.

Normally, an analysis of  $p=10^{\text{th}}$  is conducted for audio signals whose band is limited to 0.3 kHz to 3.4 kHz.

As a result of this analysis, in principle, resonant frequency of  $p/2$  spectrums, i.e., so-called a formant frequency is specified. In other words, when  $p=10$ , 5 frequencies can be specified. Then, if these 5 frequencies were allocated to both low and high frequency bands, as shown in FIG. 4, setting would be possible so that any 6 out of 16 frequencies can be covered and 2 pairs can cover any 12 frequencies.

In FIG. 4, numerals 1, 2, 3, 4, 5 refer to the allocated frequencies by

+1 cluster and I, II, III, IV, and V refer to the allocated frequencies by +1b cluster.

As PB signals, it is common that out of 16 signals, 10 digits and 2 signals (for instance, • mark and + mark) for control are used, meaning it would be OK if 12 signals could be detected. In Japan, while 4 lower frequencies (697, 770, 852 and 941 Hz) are available; only 3 higher frequencies (1209, 1336, and 1497 Hz) are available.

Parameters for detecting these can be derived from the following equation:

when specified frequencies are set to  $\{f_1\}=(f_1, f_2, f_3, f_4, f_5)$ , where  $T$  is a sampling period,  $b_1$  is a resonant bandwidth of  $f_1$ , and since in the case of a PB signal, it is provided that allowed fluctuation band of signal frequency may be  $\pm 2\%$ , about  $b_1 = f_1 \times 4\%$  may be taken.

From this, the tenth degree equation as shown below is generated

(1)

and then substituted into the following

(2)

Then, Z idempotency coefficients of the equations (1) and (2) are considered  $a_1 \dots a_{10}$ , and then  $[a_1 \dots a_{10}]$  can be determined.

A reverse spectrum coefficient to be used as a phoneme norm pattern can be determined, as shown below, as a correlation coefficient that is obtained by adding 1 made as  $a_0$  to this series of  $a$ :

Phoneme parameters stored in the +2 to +15 clusters can be determined by LPC analyzing real individual PB signals.

In addition, in reality, it is not necessary to cover all of 12 PB signals, and as one example shown in FIG. 4, since +1 cluster can specify 6 kinds and +16 clusters can specify 6 kinds, recognition experiments for the second column have only to be carried out for these 6 kinds.

We now describe one embodiment of the present invention with reference to FIG. 2.

Input voice 20 (it may be a PB signal as pseudo-voice waveform) is subjected to computation of a correlation coefficient  $\{r_1^{<a>}\}$  LPC (linear

prediction) at the voice analyzing unit 21, and residual power  $E0^x$  is computed.

Then, for every frame, by a phoneme norm pattern  $\{A1^{(a)}\}$  wherein  $I=0$  to 10 and  $n=1$  to  $S$ , a correlation coefficient  $\{r1^{(x)}\}$ , wherein  $I=0$  to 10, and  $E0$ , and by the following equation, a likelihood ratio is calculated at the distance computing unit 23:

(3)

Then, matching is taken between an input phoneme series matrix with  $L \times a$  as a scale and a phoneme symbol series word dictionary by DP matching, and one having optimum matching is output as result of recognition. In that case, as already discussed, in the 16 pairs of phoneme clusters, recognition of the first column using only 2 representative ones, e.g., +1 (male voice representative) and +16 (female voice representative) takes place, narrowing down to  $N$  candidate words. Then, when if it is determined by the pattern for PB signal detection that has been added to the +1 and +16 clusters that a first candidate is a PB signal, recognition in the second column takes place with 6 out of 12 kinds of PB signals as  $N$  candidates. Others are exactly same as the conventional voice recognition.

In this case, if a cluster for a PB signal were configured as the 17<sup>th</sup> pair without adding one each of pseudo-phoneme patterns associated with individual PB signals to 2<sup>nd</sup> pair to 15<sup>th</sup> pair, configuration could be possible

such that when the candidate is detected as a PB signal in the first column, frame-by-frame phoneme recognition has only to take place for this cluster.

In addition, in the voice recognition of more than one norm pattern at generally conducted word level, it is obvious that as for PB signals, only recognition of the first column that has been described in the 16 pair clusters, each and every one of which is subjected to voice recognition, is sufficient.

As described above, according to the present invention, voice and PB signals can be utilized as an information input means by means of telephone and without making any distinction between them, which enables input of information that takes advantage of features such as convenience of voice input and reliability of PB input.

For instance, use could be possible wherein only control words that are relatively long and can easily utilize context effects are inputted by voice, while numeric data that is short and cannot utilize context effects are inputted by PB input.

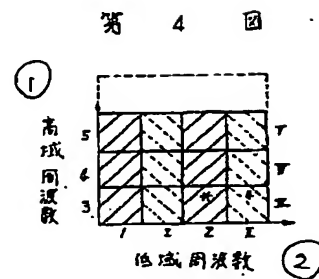
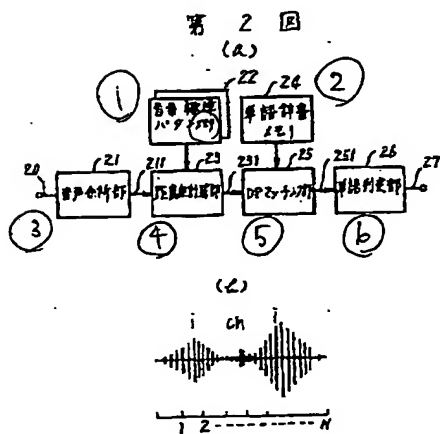
Alternatively, those who can utilize PB phones can be offered reliable PB input, while a system using voice input can service to those who cannot.

FIG. 2

- 1 Phoneme norm pattern memory
- 2 Word dictionary memory
- 3 Voice analyzing unit
- 4 Distance calculating unit
- 5 DP matching unit
- 6 Word judging unit

FIG. 4

- 1 High frequency
- 2 Low frequency



⑬ 日本国特許庁 (JP)

⑭ 特許出願公開

## ⑫ 公開特許公報 (A)

昭58—193595

⑮ Int. Cl.<sup>3</sup>

G 10 L 1/00

H 04 M 1/26

11/00

識別記号

庁内整理番号

7350—5D

7251—5K

7345—5K

⑯ 公開 昭和58年(1983)11月11日

発明の数 1

審査請求 未請求

(全 5 頁)

## ⑭ 電話情報入力装置

地株式会社日立製作所中央研究所内

⑰ 特 願 昭57—75282

⑱ 出 願 昭57(1982)5月7日

⑲ 発 明 者 中田和男

国分寺市東恋ヶ窪1丁目280番

⑳ 出 願 人 株式会社日立製作所

東京都千代田区丸の内1丁目5番1号

㉑ 代 理 人 弁理士 薄田利幸

## 明 細 書

発明の名称 電話情報入力装置

## 特許請求の範囲

1. 話者別に分類された複数組の音楽標準ボタンと認識すべき単語に対応した音楽系列単語辞書とを有する音声認識装置において、押しボタン信号の有無を検出するための第1の擬似音楽ボタンを音楽標準ボタンの特定の組の中に有し、残りの組内に個々の押ボタン信号を認識するための第2の擬似音楽ボタンを持ち、第1および第2の擬似音楽ボタンに対応した擬似音楽系列単語辞書を設けたことを特徴とする電話情報入力装置。

## 発明の詳細な説明

本発明は電話による情報の入力、とくに音声認識を利用した情報入力装置に係り、特にその機能を押ボタン信号による入力の併用にも拡大するの好適な音声認識装置の構成に関する。

従来の電話機による情報の入力手段には次の2つがある。1) 押しボタン信号入力(以下PB入

力と略す)、2) 音声認識入力(以下音声入力と略す)。

1) は音声帯域正弦波2周波(高域、低域各1周波)の組み合わせによる人工的な信号で、現在規格を統一されて使用されているものは低域4周波、高域4周波の組み合わせで原理的に16種類の情報を入力することができる(松坂、上原、矢谷: 押しボタンダイヤル電話用信号方式: 日本電信電話公社電気通信研究所研究実用化報告17-11, P241 昭和43年11月参照)。

この1)の方法によれば情報は確実に入力できるが、情報をすべて数字コードに変換して入力しなければならず、また押しボタン電話機が使えないところでは情報を入力することができない。

2) は音声認識によつて、音声のまま情報を直接入力しようとするもので、便利ではあるが、常に確実、正確に情報が入力できるとは限らない(長島、中津: 音韻単位の標準ボタンを用いた長時間単語音声認識装置、日本音響学会音声研究会資料、878-22, 1979, 渡辺、亘理、千葉

他：不特定話者用音声認識装置SR-1000シリーズ，日本音響学会講演論文集，3-1-24，1981年5月）

本発明の目的は、従来の音声認識装置の構成を基本とし、これにごくわずかの追加を行うことによつて、あらかじめ音声信号かPB信号かがわからなくても、それぞれ認識が行なわれ、そのことによつて音声とPB信号を自由に併用して使用でき、電話機による情報入力機能を拡大する手段を提供することにある。

まず、従来の電話情報入力用音声認識システムの構成を第1図に示す。

第1図において、加入者電話機11から交換機12を通つた音声信号121は音声認識部13に入力され、業務処理部14からの認識要求信号141を受けてその認識処理をおこなう。主業務処理部14では、認識結果を確認するために認識完了信号142を受けて音声出力部15に出力要求信号151を送出し、音声出力の終了を出力完了信号152により確認する。

と単語辞書メモリ24中に格納されている標準単語（たとえば、単語番号1，2，……に対応してそれぞれ音素記号系列l・chl，nl，……などで表わされる単語）との非線形マッチング演算がDPマッチング部25においておこなわれ、その結果得られた距離和251の大きさにもとづいて単語判定部26で入力音声の判定がおこなわれ、認識結果27が出力される。

この認識処理の特徴は、電話入力された不特定話者の音声認識を、16組の音素標準パターンによるフレーム別認識をおこなう第1段と、フレーム別認識の結果と音素記号系列単語辞書とのDPマッチングをおこなう第2段とからなる2段のパターン整合に分解し、第1段では音素標準パターンにたいして話者の音声波形における音響的な特性にもとづいて16組のクラスタリング（組み分け）をおこない、第2段では1つの単語に対して複數個の音素記号系列単語辞書をもうけて、発話の変化、たとえば母音の無声化や鼻音化、に対処していることである。本方式はこの2段処理によつて

一方、交換機12からの応答信号122を受けて発信制御部16から出力された応答信号161が主業務処理部14に入力されると、電文処理部17にたいして送信要求信号143が送出される。

これを受けた電文処理部17は通信制御部18にたいし、送信要求信号171を送ることによりリレーコンピュータ19から発せられ通信制御部18を通つた電文181を受信して発信制御部16にたいし発信要求信号172を送り信号162を発信させる。

第2図は第1図における音声認識部13のブロック構成を示す。

第2図(b)で示す波形の入力音声20（ichi）から音声分析部21において抽出された特徴パラメータの系列211と音素標準パターンメモリ22中に格納されている例えば16組の音素（a，i，……など）の特徴パラメータ（最尤スペクトルパラメータ、LPCケプストラム係数など）との距離が距離計算部23において計算される。

距離計算部23から出力された距離の系列231

所要メモリ量が少なくかつ不特定話者の音声に対して高い認識能力を持つことが知られている。

「一桁の数字音（0～9の10語）」および、「はい」、「いいえ」、「どうぞ」、「もう一度」、「ほりゆう（保留）」、「とりけし（取り消し）」の6語を含む16語に対して620名の男女による認識結果の一例を表1に示す（電電公社通信研究所発表）。

表1 男女別の誤り率〔%〕

	男	女	平均
尤度による距離	4.45	3.22	3.86
LPCケプストラム距離	4.57	3.18	3.90

なお、表1で距離尺度としてとられているのは、音声認識のための特徴として使われるパラメータの一例であり、このいづれを用いても誤り率はほとんど差のないことをあらわしている。

この方式のもう一つの特徴は、16組に分類された多数の（最大40個程度）音素標準パターンとの整合によつて、フレーム別に音素系列を認識

し、その結果と単語音素系列との比較によつて単語を認識するに当つて、その処理量を軽減し、実時間認識を可能にするため、そのフレーム別音素認識を第3図(a)に示すように2段に分けて階層的に行っていることである。すなわち、16組のパタンの中、男声の代表として作られている例えば第1の組と、女声の代表として作られている例えば第16の組との2組の標準パターンで、まず第1段の認識を行い、その中で整合度の良いものN語をえらび、そのN語に対象を限定し、改めて上記16組の音素標準パタンのすべてを使つて再認識を行う。Nの数としては第3図(b)に示す実験結果から $N=4$ にとれば、誤認識による誤りが少なく、処理量(計算量)も少なくて済むことがわかる。ここで計算量の比率とは、

$$\frac{\text{一次選択ありの計算量}}{\text{一次選択なしの計算量}} \times 100 \text{ で与えられる。}$$

結果的に16語に対して16語 $\times$ 16語 $=256$ 組 $\times$ 語の処理を、2組 $\times$ 16語 $+16$ 組 $\times$ 4語 $=96$ 組 $\times$ 語の処理に軽減している。

応する単語辞書としては、受信検出しなければならないとされている時間以上(たとえば現行規定によれば40ミリ秒以上)同一のPB信号に対する擬音素標準が維持するという条件を満足するように構成すればよい。

次に最初2個の代表クラスターによつて第1段目の認識が行なわれるという階層処理の場合には次のように考える。

1) 第1段目の認識でPB信号であることを検出する。

2) 第1段目認識でPB信号と検出された場合第2段目でそのいずれであるかを認識する。

以下さらに具体的に説明する。

音韻認識において、LPC(線形予測)分析にもとづいて尤度比による整合をとる場合について考える。

0.3 kHzから3.4 kHzまでに帯域制限された音声信号に対して、通常 $p=10$ 次の分析が行なわれる。

この分析の結果、原理的には $p/2$ 個のスペク

さて通常のPB信号は、いわゆるPB信号受信器で検出される。通常の使い方は情報を入力する信号の形式がPB信号であるか音声信号であるかはあらかじめ決まっており、分離して行なわれる。例えば通常の使い方はPB信号は情報センタへのアプローチに使われ、第1図における発信制御部13で受信検出される。

本発明のポイントはすでに述べた第1図の音声認識部13へ、音素および単語の1個としてPB信号を加え、音声認識と全く同じ形式でPB信号を検出することによつて、音声とPB信号を情報入力手段として併用してもよいようにしようとするものである。ただし音声信号とPB信号の同時共存は仮定しない。

まず簡単に、16組の音素パタンのすべてと総当たりでフレーム別音素認識が行なわれる場合を考える。このときは16組の各組に1個、あわせて16種のPB信号を擬音素標準として割り当て、その検出に必要な特徴パターンを音素標準パターンメモリに記憶させておけばよい。PB信号に対

トルの共振周波数いわゆるホルマント周波数が指定される。すなわち $p=10$ の場合、5個の周波数を指定することができる。この5個の周波数を、低、高の両周波数帯に、第4図に示すように割り当てれば、16個の周波数の中の任意の6個をカバーするように設定することができ、2組によつて任意の12個をカバーするようにすることができる。

第4図において、1, 2, 3, 4, 5は $\phi 1$ クラスターの割り当て周波数を示し、I, II, III, IVは $\phi 1b$ クラスターの割り当て周波数を示す。

PB信号としては、16個の中から実際には10数字と制御用に2個(たとえば $\phi$ 印と $\phi$ 印)が用いられるのが普通であり、12個を検出できればよい。日本国内では低域は4周波(697, 770, 852, 941 Hz)であるが、高域は3周波(1209, 1336, 1497 Hz)しか使っていない。

これらを検出するパラメータは次式から導出することができる。

指定周波数を  $\{f_i\} = (f_1, f_2, f_3, \dots, f_i, f_{i+1}, \dots)$  とするとき

$$\begin{cases} \beta_i = r_i e^{j\theta_i} & r_i = e^{-\pi b_i T} \\ \bar{\beta}_i = r_i e^{-j\theta_i} & \theta_i = 2\pi f_i T \end{cases}$$

ここで  $T$  はサンプリング周期、 $b_i$  は  $f_i$  の共振帯域幅であり、PB信号の場合、許容信号周波数変動幅は  $\pm 2\%$  と規定されているから  $b_i = f_i \times 4\%$  程度にとればよい。

これから

$$\begin{aligned} (Z - \beta_1)(Z - \bar{\beta}_1)(Z - \beta_2)(Z - \bar{\beta}_2) \\ (Z - \beta_3)(Z - \bar{\beta}_3) \dots \dots \dots (1) \end{aligned}$$

の10次方程式を作り、それを

$$Z^{10} + \alpha_1 Z^9 + \alpha_2 Z^8 + \dots + \alpha_9 Z + \alpha_{10} \dots \dots \dots (2)$$

とわいて(1)式と(2)式の  $Z$  の等べき係数を  $\alpha_1, \dots, \alpha_{10}$  とおけば、 $(\alpha_1, \dots, \alpha_{10})$  が求められる。

音韻標準パターンとして使われる逆スペクトル係

れ、残差電力  $E_s$  が計算される。

次に距離計算部23で各フレーム毎に音楽標準パターン  $\{A_i^{(n)}\}$ ,  $i = 0 \sim 10$ ,  $n = 1 \sim 8$  と入力  $X$  の相関係数  $\{r_i^{(n)}\}$ ,  $i = 0 \sim 10$  と  $E_s$  によつて次式によつて尤度比が計算される。

$$L_s = (A_0 \cdot r_0 + 2 \sum_{i=1}^{10} A_i \cdot r_i) / E_s \dots \dots \dots (3)$$

この  $L_s$  を尺度とする入力音楽系列マトリックスと音楽記号系列単語辞書との間でDPマッチングによる整合がとられ、最適整合のものが認識結果として出力される。その場合、すでに説明したように16組の音韻クラスにおいて、代表的な2つ、たとえば $\phi 1$  (男声代表)と $\phi 16$  (女声代表)のみを用いた第1段の認識が行われ、候補単語が  $N$  個にしぼられる。このとき、 $\phi 1$  と  $\phi 16$  のクラスターに追加されたPB信号検出用のパターンによつて第1候補がPB信号であると検出されたときは、 $N$  個の候補として、12種類のPB信号の中の6個を候補として第2段目の認識

数は、この  $\alpha$  の系列に  $\alpha_0$  とした1を加えた系列の相関係数として、

$$\begin{aligned} A_0 &= 1 + \alpha_1^2 + \alpha_2^2 + \dots + \alpha_{10}^2 \\ A_1 &= \alpha_1 + \alpha_1 \alpha_2 + \alpha_2 \alpha_3 + \dots + \alpha_9 \alpha_{10} \\ &\vdots \\ A_9 &= \alpha_9 + \alpha_9 \alpha_{10} \\ A_{10} &= \alpha_{10} \end{aligned}$$

と求められる。

$\phi 2$  から  $\phi 15$  までのクラスターに記憶される音楽パラメータは、現実の個々のPB信号をLPC分析することによつて求めることができる。

なお実際には12個のPB信号をすべて対象とする必要はなく、第4図にその1例を示すように、 $\phi 1$  クラスターによつて6種類、 $\phi 16$  クラスターによつて6種類が指定されるから、この6種類についてのみ第2段の認識実験を行えばよい。

本発明の一実施例を第2図を用いて説明する。

入力音声20 (擬似音声波形としてPB信号であることもある) は音声分析部21で相関係数  $\{r_i^{(n)}\}$  の算出とLPC (波形予測) 分析がさ

を行ふ。その他は従来の音声認識と全く同じである。

この場合、個別PB信号に対応する擬似音楽パターンを2組から15組に1個ずつ加えないで、第17組としてPB信号用のクラスターを構成すれば、第一段目でPB信号として検出されたときは、このクラスターについてのみフレーム別音楽認識を行えばよいように構成することもできる。

また一般に行なわれている単語レベルでの複数標準パターンによる音声認識においては、PB信号に対しては16組クラスター総当りで説明した一段目の認識のみでよいことは自明である。

以上説明したように本発明によれば、音声とPB信号を何ら区別することなく電話による情報入力手段として利用することができ、音声入力の簡便さとPB入力の確実さの特色を活かした情報入力が可能となる。

たとえば、音声によつては比較的長く、文脈効果の利用しやすい制御語のみを入力し、短かくて文脈効果の利用しえない数字データはPB入力と

するといった使い方も可能となる。

あるいはPB電話機を利用できる人には確実なPB入力を、利用できない人には音声入力を使うシステムをサービスすることもできる。

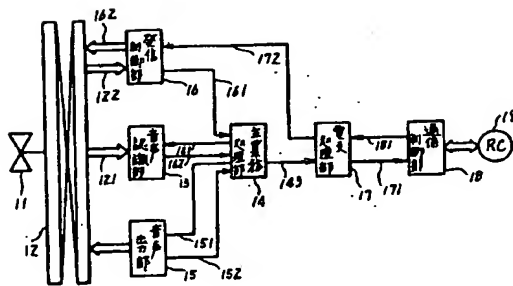
図面の簡単な説明

第1図は従来音声認識応答システムの構成図、第2図はその音声認識部の説明図、第3図は実際に行なわれている階層認識処理の説明図、第4図はPB信号検出用擬似音韻パターンによる検出可能領域の説明図である。

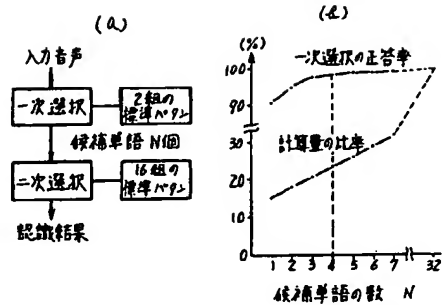
13…音声認識部。

代理人 弁理士 薄田利幸

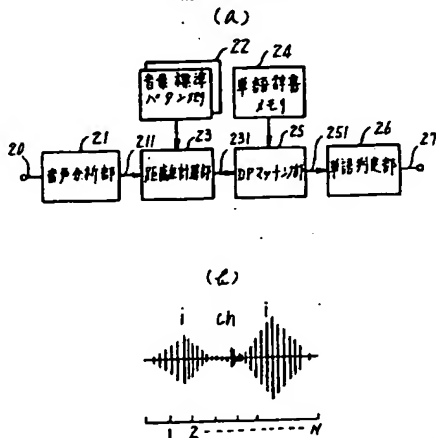
第1図



第3図



第2図



第4図

